International Journal of Contemporary Educational Research (IJCER)

www.ijcer.net

Developing a Macroscopic Lens Into Middle School Reform: Psychometric Properties of the AMLE SIA

Ayse Tugba Oner[1], **Bilgin Navruz**[1], **and Robert M. Capraro**[1]
[1]Texas A&M University

To cite this article:

Oner, A.T., Navruz, B., & Capraro, R.M. (2015). Developing a macroscopic lens ınto middle school reform: psychometric properties of the AMLE SIA. *International Journal of Contemporary Educational Research, 2*(2), 89-103.

IJCER International Journal of Contemporary Educational Research
Volume 2, Number 2, July 2015, Page 89-103 ISSN: 2148-3868

Developing a Macroscopic Lens Into Middle School Reform: Psychometric Properties of the AMLE SIA

Ayse Tugba Oner[1]*, Bilgin Navruz, Robert M. Capraro[1]
[1]Texas A&M University

Abstract

The purpose of this study was to assess the validity and reliability of the AMLE SIA, which was developed by the NMSA to provide the best educational programs for young adolescents to improve their skills. To promote these skills, NMSA suggested that schools needs to implement 16 characteristics nested within three categories. However, many middle schools failed to implement the practices. The reason might be the instrument itself due to including 96 items and the design. Therefore, the validity of the instrument was analyzed; response organization system was redesigned; the items were revised and eliminated by using regression (83 items); and the final instrument's validity was analyzed by using EFA (73% of variance explained) and the reliability (.98) was calculated.

Key words: School Improvement Assessment, Validity, Reliability

Introduction

Young adults' future life is likely to be affected by decisions made between the ages of 10 and 15. Students decide about next steps during middle-level school years and school environment is an important component in their lifetime decision. Therefore, educational programs for young adolescents need to represent the best environment for 10-to 15- year- olds (National Middle School Association, 2010). In middle school, best practices have a positive effect on student achievement (Cook, Faulkner, & Kinne, 2009; Jackson & Lunenburg, 2010; McEwin & Greene, 2011; Mertens & Flowers, 2006). However, student achievement should not be the only concern; educators need to be aware of other aspects. Middle level education occurs during crucial years in students' educational lives because during these "transitional years" (National Middle School Association, 2010) students are likely to undergo significant changes in many aspects: physical, intellectual, moral, psychological, and socio-emotional. Any unwanted situation could affect these aspects, which could result in poor high school performance and high school dropout (National Middle School Association, 2010), which educators would not like to encounter. Thus, young adults needed an education that would improve their aspects and lead them to be optimistic about the future (National Middle School Association, 2014).

Students need an education that prepares them to overcome the present century's difficulties. Success in the ever-changing world is one factor that needs to be addressed in good education (National Middle School Association, 2014). The changing world causes changes in education, and educators must continue developing and trying to sustain the success of middle grade schools. During this development, educators use the attributes and characteristics recommended in *This We Believe: Keys to Educating Young Adolescents* (National Middle School Association, 2010), in which the National Middle School Association (NMSA) asserted knowledge and many skills that students should have to be self-actualized, fully functioning people. These skills and knowledge were major goals of middle level educators. For instance, students need to think critically and rationally; be able to gather, assess, and interpret data deeply; develop skills and interests; use digital tools to gather information from different sources; take responsibilities and make ethical decisions about their own health and wellness. To achieve these goals, educational programs should be developmentally responsive, challenging, empowering, and equitable (National Middle School Association, 2010), so students can overcome difficulties.

To achieve these goals, educational programs need to be aligned with 16 characteristics that were asserted by NMSA. The 16 characteristics were grouped into three categories (National Middle School Association, 2014):

* Corresponding Author: *Ayse Tugba Oner, aysetugbaoner@tamu.edu*

1) Curriculum, Instruction, and Assessment; 2) Leadership and Organization; and 3) Culture and Community. The Curriculum, Instruction, and Assessment Category dealt with having an engaged, active, purposeful, challenging, integrated learning and teaching environment and assessment. The Leadership and Organization Category concerned supporting professional development, collaboration, and organization for all stakeholders. The Culture and Community Category dealt with giving support and guidance to students by having a safe and inviting school environment. Having practices aligned with these three categories was useful for schools in terms of reintroducing successful and promising young adults to society. For instance, research conducted in a middle school in Kentucky showed that if a middle school concept aligned with the guidelines provided by *This We Believe: Keys to Educating Young Adolescents*, the students' academic achievement was higher than that of students in schools not aligned with the guidelines (Cook et al., 2009).

To assess whether a school's program aligns with these categories, NMSA introduced an instrument, the Association for Middle Level Education (AMLE) School Improvement Assessment (SIA), for middle school educators to evaluate middle schools. The AMLE SIA was based on 16 characteristics of *This We Believe: Keys to Educating Young Adolescents.* The SIA also provided reports about strengths and weaknesses of schools. By knowing the strengths and weaknesses of schools, teachers had a chance to help young adolescents become successful and responsible global citizens that were aimed in *This We Believe: Keys to Educating Young Adolescents.*

Even if *This We Believe: Keys to Educating Young Adolescents* presented the best practices in middle-level education, many middle level schools failed to fully implement middle-level practices (McEwin & Greene, 2011). One possible reason for this is that the instrument prepared by NMSA might not provide sufficient information about schools and their needs, which could be the result of an assessment that does not measure what it intend to assess. Therefore, the investigation of the SIA could be helpful in providing a valid instrument scores for evaluating both best practices in middle-level education and schools in terms of categories provided by NMSA (2011).

Assessing the assessment itself could be the first step before using that assessment. There are important questions that need to be answered in research, such as whether the assessment's content measures the intended purpose of the assessment and includes all related contents within it, how well the assessment predicts the criterion, and whether the measure assess the construct that is intended to represent. These questions deal with the validity of an assessment. Validity focuses on "whether a particular inference or conclusion is correct, reasonable, or accurate" (Bryant, 2000, p.101). Nunnally (1967) explicitly defined construct validity - one type of validity - as

> The degree to which it is necessary and difficult to validate measures of psychological variables is proportional to the degree to which the variable is concrete and abstract….To the extent that a variable is abstract rather than concrete, we speak of it as being a construct. Such a variable is…something that does not exist as an isolated, observable dimension of behavior. (pp. 84-85)

Therefore, it is important to have a measurement that can be accountable in terms of validity of constructs.

The purpose of this research was to analyze the validity and reliability of the Association for Middle Level Education (AMLE) School Improvement Assessment (SIA) and improve the organization of the SIA for future application. The reason was that there were concerns about the instrument, such as items in the assessment did not align well with the 16 Characteristics, or it was wordy and subject to spurious interpretation. Also the AMLE SIA instrument had 96 items; evaluating 96 items would be time consuming. As such, we also analyzed how effective the response organization system was and what would be the most efficient way to answer items in this instrument. By re-evaluating the instrument, we expected to have an instrument better aligned with the 16 characteristics, more trustworthy and valuable information about middle-grade practices, and a more focused instrument that would enable more targeted professional development and focused school improvement efforts.

Method

The study consisted of four phases. The two of the phases were designed to provide insights into the validity of the scores obtained from the instrument, and other two of the phases focused on the instrument's functioning and sample results.

Phase 1

During Phase 1, several forms of validity were considered. Initially, Face Validity (Mosier, 1947) was used. With face validity, we checked to see if the operationalization, on its face, was a satisfactory translation of the construct we intended to measure. The intended construct was the "*This We Believe. . .*" document (National Middle School Association, 2010), which describes three categories comprised of 16 characteristics across 96 items. The 96 items were provided to 25 experts who examined the items and compared them to the statements in *This We Believe*. However, face validity is likely the weakest way to demonstrate construct validity (cf. Cronbach, & Meehl, 1955) and probably the most abused. Therefore, we took a very systematic and clinical approach. Experts were selected from the field of middle school were well versed in the constructs expressed in *This We Believe*. All the experts either were original authors of the document, reviewers of the document, or on the research advisory board for the National Middle School Associations, (has since changed its name to Association of Middle Level Educators [AMLE]).

Face validity. Face validity concerns the face of the instrument. If the instrument seems like a good measurement for the intended purpose, then it would have face validity. For instance, if an instrument is designed to measure math ability, as a researcher you might look at items and think that these items fit the instruments purpose. However, even if items are related to measuring math ability, because this is a subjective decision this type of validity might be considered as weak. The weakness of evidence does not mean that the instrument does not measure correctly; on the contrary, it might, but it still would be subjective. To prevent subjectivity, the researcher could send the instrument to other experts to improve its face validity. Therefore, besides face validity there are other validity types that should be examined to ensure the validity of an instrument.

Content validity. Content validity deals with whether an instrument evaluates relevant aspects of it (Bryant, 2000). The important issue for content validity is itself, because the test might measure a content related with the intended content which makes the test fail to have the content validity. The test also should measure all components of the content. For instance, a final exam for a course needs to measure the concepts that were taught in that course. Content validity is crucial because for some constructs, such as our final exam example, it might be easy to determine the criteria that fit to that construct, but for other constructs (e.g., attitude, intelligence) it is not (Trochim, 2006). The criterion in content validity is the measure itself. Therefore, there cannot be a correlation between a new measure and the criterion (Nunnally, 1967).

Predictive validity. The purpose of predictive validity is to "estimate some important form of behavior" (Nunnally, 1967), which is criterion. A test needs to be able to assess how successful it is in predicting the criterion before measuring it (Bryant, 2000). In that case, the instrument, theoretically, needs to predict the criterion. If there is a high correlation between predictor and criterion, the measure can predict something well. For instance, if the predictive validity of a measurement used to predict how well students perform in a specific grade is good, it should be able to predict students' success in that specific grade, not something else.

Concurrent Validity. In concurrent validity, unlike predictive validity, the scores are obtained from a new measure and criterion measure at the same time (Bryant, 2000). As in predictive validity, in concurrent validity, there should be a high correlation between these two measures. Thus, one test can be substituted for the other one because one measurement would be expected to measure the same construct. For instance, if there is a strong correlation between students' course grade and the passing test, the course-passing test could be given to students instead of the course itself. To have strong concurrent validity for an instrument, it would be better to collect test scores and criterion measure separately (Bryant, 2000).

Phase 2

Response organization redesign. It is important to pay as much attention to the online structure of the instrument as to the psychometric structure of the instrument. The instrument was originally designed to be administered one item at a time nested within characteristic, with characteristics nested within categories. This model required participants to be aware and to be presented with repetitious text about the category and characteristic to which they were responding. To examine the online structure of the instrument we considered the time it took to complete the instrument, time spent per item, and ways to improve the response time. Initially, we plotted the average item response time. For this part, there were 24 middle grades teachers, staff, and administrators. Then, we randomized the items to determine if there was any pattern with regard to the order of the characteristics and categories. A new sample from a different middle school was selected (n= 22). The reason for the randomization of items was to rule out whether the average response time was a factor of the item's difficulty toward the end of the test or due to the length. Finally, we revised the online layout and item response format again examining average item response time. Items were again nested within characteristics and characteristics nested within categories; however, now all of the items for a characteristic were displayed at the same time within the same block, and sliders were used for input. This effectively reduced the number of screens to 16 and eliminated the need to scroll. For this portion, a new sample of middle grades teachers, staff, and administrators from another middle school were selected (n= 23). The total number of items examined was 96.

Phase 3

This phase consists of two parts, a 15-member panel of experts and a separate group of six experts to review the comments and ratings for the purpose of deleting or revising items. In this phase, to have a more robust instrument, we analyzed it according to responses and comments from 15 middle grades experts. All experts were either classroom teachers (6) with 15 or more years of teaching and participation in AMLE or university middle level teacher educators (9) with 20 or more years combined of experience teaching middle school and university experiences researching middle level. They rated the items for relevance to the category and to the characteristic. Participants rated the items on a 100-point relevance rating scale where 1 indicated that the item did not align with the characteristic and category and 100 indicated that the item aligned well. They also were asked to provide qualitative commentary for each item. The qualitative commentary they were asked to cover concerned whether the item was relevant, possible rewording options, redundancy, or should be removed. For each item rated 50 or below, a new box would appear asking for a revision of the item or if the item should be removed from the scale and a reason.

Primarily, in the original AMLE SIA, there were 96 items. The first Category, Curriculum, Instruction, and Assessment, included 5 characteristics with 39 items. The Leadership and Organization Category consisted of 5 characteristics with 27 items. The last Category, Culture and Community, was comprised of 6 characteristics with 30 items.

All scores were converted to a percentage scale to make interpretation of data easier. Data were analyzed by aggregating scores across raters and computing means and standard deviations. Regression was used, and standardized β weights were obtained with the total characteristic score as dependent and weighted item ratings as predictors. Structure coefficients were computed by correlating predicted values with the dependent variable. The structure coefficient estimates the percentage of useful variance accounted for by each item independent of other items. We used this as a relative measure of the importance of each item without considering whether or not that variance accounted for estimate was unique.

A separate and distinct group of six experts met to make final decisions for item revision or deletion. After consideration of quantitative and qualitative information, items with a rating of 90 or above were decided as adequate and retained; items with a rating between 70 and 90 were considered strong candidates for revision; and items with less than 70 were considered strong candidates for removal. In addition, if two items had equivalent ratings and the qualitative information indicated that the two items were addressing the same information, the item that had the higher β weight was considered for revision whereas the other item was deleted. As a result, for the first Category, seven items were removed, eight items were selected for revision, and 24 items performed well. Three items from the second Category and three items from the third were eliminated. In the second Category, five and 19 items were considered for revision and retention, respectively. Seven items were selected for revision and 20 items performed well for the last Category.

Phase 4

EFA analysis. The sample included 15 teachers, nine of which were female. Seven of 15 individuals completed the survey by answering all the items.The final form of AMLE SIA included 83 items and was administered on the Web. Participants responded to items using a 0-to-100 unnumbered graphic rating scale. A 0 indicated never and 100 indicated all the time. Users moved a slider to indicate their response in the Qualtrics online survey software. The answer could be any number on the 0-100 scale, such as 47, 78, or 99. We chose the unnumbered graphic rating scale because it has been shown to be a favorable method for collecting psychometrically reliable scores (Cook, Heath, Thompson, & Thompson, 2001) (see Figure 4).

In the data set ~14% was missing observations, so as the first step for analysis was to impute data using the expectation-maximization (EM) algorithm, which is one of the popular methods for dealing with missing data. The Proc MI procedure in SAS 9.3 was used to impute the data. The Proc MI procedure reads raw data with missing observations and provides maximum likelihood variance-covariance estimates with the vector of means.

The EM covariance matrix is an excellent approach to deal with missing data prior to exploratory factor analysis (EFA) applications (Graham, 2012).

EFA was conducted to examine construct validity of scores. In EFA studies, the desired sample size was suggested as more than 200 (e.g., Cattell, 1978; Guilford, 1954). However, there were simulation studies that showed that EFA also provides sufficient results with a small sample size between 10 and 50 (Mundfrom, Shaw, & Ke, 2005; Preacher & MacCallum, 2002). Recently, it was found that EFA yielded reliable results even when sample size was smaller than 50 (i.e. even smaller than 10 in well conditioned data) (de Winter, Dodou, & Wieringa, 2009). Therefore, even in this study, in which there were 15 cases, conducting EFA would not yield misleading results.

After we obtained our variance-covariance matrix, we conducted most of the parametric statistical methods by using the information provided in that matrix, rather than using raw data (Zientek & Thomson, 2009). When we use sufficient decimal places in the variance covariance matrix, or in the correlation matrix with standard deviations of those variables, we would exactly estimate the same results, which can be found by reading raw data. Thus, we can use EM estimated variance covariance matrix in our EFA analysis to estimate our parameters of interest because EFA is one of the parametric multivariate models under the general linear model (GLM).

Principal components analysis was used as our factor extraction method. There were other methods to extract factors, such as principal axis factoring or maximum likelihood. Robust studies with smaller samples never used maximum likelihood as their method of choice. In terms of comparison of principal component and principal axis extraction methods, the two methods yield equivalent results when variables are increasing in the study or variables are reliable (Authors, 2004; Thompson, 1992). There were 83 items in the instrument, so we believe that either of these two choices would have returned the same results.

In EFA studies, the next decision is to determine the number of factors to extract from the correlation matrix. There are several strategies to decide how many factors should be extracted, such as Kaiser's eigenvalue greater than one rule, Catter's scree plot, and parallel analysis (Authors, 2004). The most commonly used method is the eigenvalue greater than one rule because this rule is default in many statistical packages (e.g., SPSS) (Thompson & Daniel, 1997). In our EFAs, we specified the number of factors as three because the instrument was based on middle grades educational theory.

Generally speaking, in EFA, after extracting the factors, rotation is almost always necessary to interpret factors easily (Thompson, 2004). There are two extraction methods: orthogonal and oblique. The difference between these rotation methods is that, philosophically, if the researcher believes the factors to be uncorrelated the orthogonal rotation retains uncorrelated characteristic in factors. For the oblique rotation the researcher must believe that the factors are correlated and the rotation retains that characteristic. Both rotation strategies are designed to obtain simple structure (Thurstone, 1947) for easier interpretation.

Correlated factors are one indicator for the existence of higher order factors (Gorsuch, 1983). While first order factors are extracted from the variable correlation matrix, higher order factors are extracted from the inter-factor correlation matrix (Thomson, 2004). Gorsuch (1983) explained higher order factors:

> Rotating obliquely in factor analysis implies that the factors do overlap and that there are, therefore, broader areas of generalizability than just a primary factor. Implicit in all oblique rotations are higher-order factors. It is recommended that these be extracted and examined so that the investigator may gain the fullest possible understanding of the data. (p. 255)

This portion of the study included two parts. In the first part, the authors reviewed the items, and they expected three correlated factors would be extracted under one higher order factor. We extracted three first-order factors, and one higher order factor was extracted from the "three by three by three" inter factor correlation matrix. Promax rotation, one of the oblique rotation strategies, was used to rotate three first order factors for easier interpretation. In the second part, we conducted three separate EFAs for each of the three categories separately. In each EFA, we extracted one factor from the inter-variable correlation matrix by using principal component analysis.

Results

The results section moves from the evidence of score validity, examination of the response method to item revision, and finally exploratory factor analysis of the revised instrument.

Results of Phase 1: Validity

The expert review of face validity was in agreement that the items reflected the constructs expressed in the *This We Believe . . .* document. The rating was 100% agreement for the items in Characteristics 1, 6, 10, 11, 13, 15, 16, and 92% agreement for the items in Characteristics 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14. However, there was more variability in Characteristics within categories. For characteristics within category there was 88% agreement for Category 1, 91% agreement for Category 2, and 84% for Category 3. The major conflict in agreement when considering the higher-level construct was that some characteristics could belong to two of the categories depending on the interpretations of the people answering the items. Primarily, the expert conversations were dominated by concerns for individual items comprising characteristics that could lead to the erroneous interpretations. Therefore, the problematic items were either removed or revised depending on consensus. After alterations and deletions, the group had 100% agreement across all items, characteristics, and categories.

For the concurrent validity we assessed the ability of the instrument to distinguish between those who knew the principles of *This We Believe . . .* from those who were unfamiliar. We compared the responses from the expert panel to those of 16 high school teachers who had never been middle school teachers, were not members of the Association of Middle Level Educators, and had not read the *This We Believe . . .* document. The two groups responded to the items and chose a 1 or 0, 1 indicating it was contained in the *This We Believe . . .* document and a 0 indicating that it was not. The score discrepancy was stark. Scores for the expert panel were very high (mean = 4.71, SD= .43) while the high school teachers were lower (mean = 3.15, SD= .99). Some potential reasons for why the score discrepancy was not even greater could lie in the fact that some of the principles are closely tied to good educational principles that cannot be isolated to the middle school only. However, when the groups were asked to determine if the items were contained in the *This We Believe . . .* document the difference was more dramatic. The chance score for the rating was .50 (there were only two choices, either a 0 or a 1, so participants had a 50/50 chance of guessing it correctly. The par score was .92, or 92 percent of the items were part of the document. The high school teachers' rating was .54, just above chance but well below the par score. The experts' score was .88. While they were able to identify all the distractors with 100% accuracy, they misclassified some items intended to measure the *This We Believe . . .* document. The identified items were the items that were flagged earlier as potentially problematic and subsequently either dropped or revised. The magnitude of the effect was .34, a sizeable difference between the groups.

Results of Phase 2: Response Organization Redesign

According to the results from 24 participants for items grouped within characteristic and characteristics within categories, participants' mean administration time was 34 minutes and 20 seconds. Participants' item response time grew longer toward the end of the assessment (see Figure 1). The assessment's items were not longer toward the end nor were they more complex; rather, respondents might suffer fatigue because of the total number of items (96) or the presentation of the items. As such, their response time was unacceptably long (see Figure 1) and calls into question the dependability of their responses toward the end of the administration.

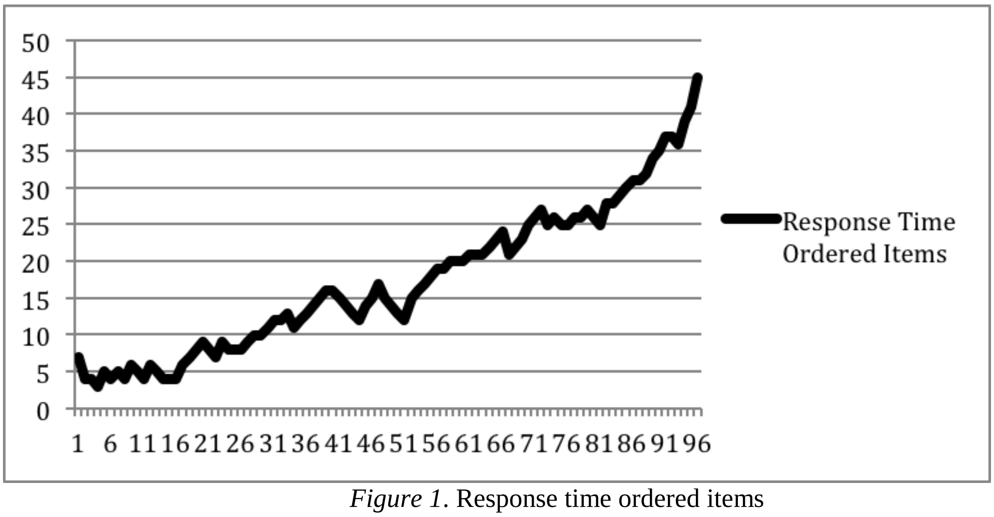


Figure 1. Response time ordered items

The results for the second type of response organization, random ordered items (not grouped by category or characteristic), showed that it was harder to keep the same response time performance when items were organized randomly. To complete the item the respondent was required to reach the descriptor for the category and the characteristic for every item. The mean administration time was 35 minutes and 44 seconds. The response time of participants was longer than the first organization type (see Figure 2). Organizing items randomly might create a chaotic environment for respondents. When two figures of two types of organization were taken into consideration (Figure 1 and Figure 2), the random ordered organization system graphic appeared more jagged in comparison to ordered items' graphic.

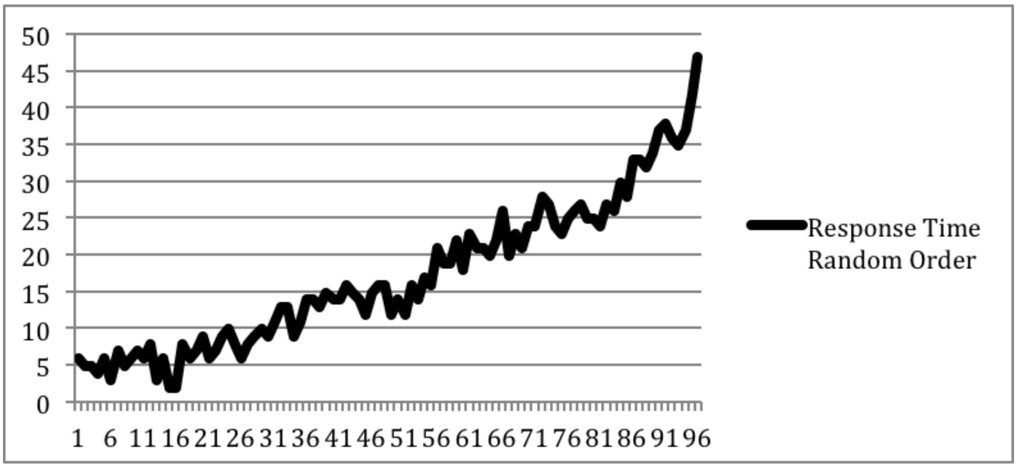


Figure 2. Response time random ordered items by standard presentation

The result for the systematic slider organization, items grouped by characteristic within category by slider, was the best one for respondents' (see Figure 3). The minimum and maximum response times per item were 3 and 13 minutes, respectively. The average administration time was 8 minutes.

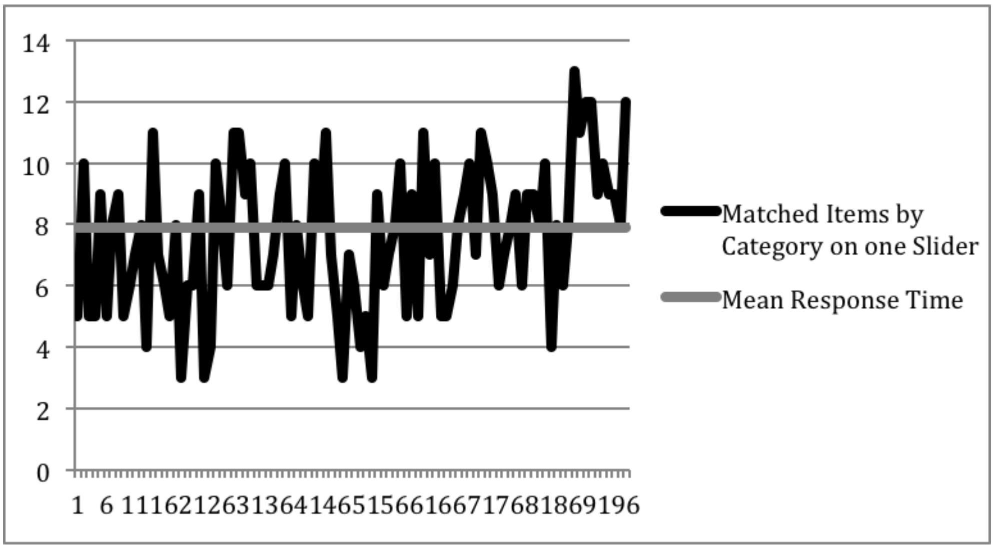
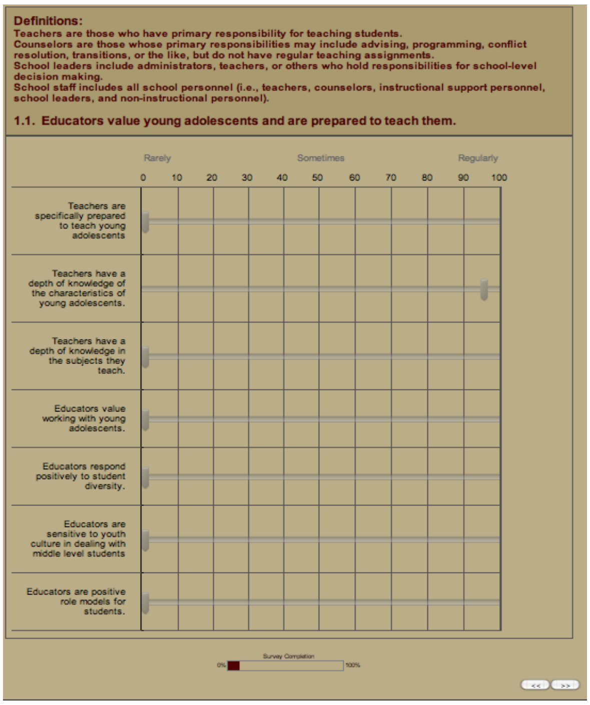


Figure 3. Response time ordered items grouped by slider choice

In this organization, the direction appeared only once per characteristic (see Figure 4). Also, this organization moved to the idea of 100-point scale, which most school personnel were very accustomed. The progress bar at the bottom of the electronic response page divided into 16 segments as opposed to 96. According to these results, it is important to organize items into one category, and present these items together in an online page in order to help respondents. Formatting the assessment plays a crucial role for the target group. It is important to notice that the variation around the mean was consistent and there was not quadratic appearance to the response time as the test time progressed. In addition, the average completion time was decreased to one-fourth of the time necessary.



Definitions:
Teachers are those who have primary responsibility for teaching students.
Counselors are those whose primary responsibilities may include advising, programming, conflict resolution, transitions, or the like, but do not have regular teaching assignments.
School leaders include administrators, teachers, or others who hold responsibilities for school-level decision making.
School staff includes all school personnel (i.e., teachers, counselors, instructional support personnel, school leaders, and non-instructional personnel).

1.1. Educators value young adolescents and are prepared to teach them.

Teachers are specifically prepared to teach young adolescents

Teachers have a depth of knowledge of the characteristics of young adolescents.

Teachers have a depth of knowledge in the subjects they teach.

Educators value working with young adolescents.

Educators respond positively to student diversity.

Educators are sensitive to youth culture in dealing with middle level students

Educators are positive role models for students.

Figure 4. Example of items grouped by slider choice

Results of Phase 4: EFA Analyses

One-first-order factor solution

Score reliability. The reliability of scores was computed by Cronbach’s α internal consistency coefficient. The reliability coefficient of whole test scores was 0.983. The internal consistency coefficient statistically showed the proportion of true score variance, so closer to 1 is desired. Our estimated reliability coefficient, 0.983, was very close to 1, thus we could say that scores were reliable.
Factor structure. To evaluate the test score validity, principal component analysis was conducted, following the guidelines described in Thompson and Daniel (1997).

Three first order factor solution

*Factor structure***.** The three eigenvalues prior to rotation for the factors subsequently were 44.861, 8.098, and 7.521, respectively. These three factors explained around 73% of the total variance of the variable correlation matrix. Because the authors expected the existence of one second-order factor, promax rotation for the simple structure was applied. Second order factors can only be extracted from an inter-factor correlation matrix after one of the oblique rotation strategies is used (Gorsuch, 1983; Thomspon, 2004). The eigenvalues for the components were, respectively, 26.322, 22.382, and 11.775 after rotation, and as expected they explained exactly the same amount of variance, 73%, of the variable correlation matrix. Table 1 shows pattern coefficients after Promax rotation was applied.

When one of the oblique rotations was applied to obtain simple structure, structure coefficients were not equal to pattern coefficients anymore, which is the case in orthogonal rotation (Thompson, 2004). Thus, whenever an oblique rotation is applied, structure coefficients should be reported separately (Thompson, 2004). Table 2 shows the structure coefficients.

Table 1. *Promax Rotated Pattern Coefficients*

Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3
1	**0.620**	0.100	0.259	22	**0.513**	0.543	-0.111	43	**0.480**	0.398	0.110	64	0.477	-0.025	**0.613**
2	-0.032	**0.454**	0.029	23	**0.508**	0.479	0.013	44	0.327	0.280	**0.548**	65	**0.688**	0.366	-0.039
3	0.160	0.369	-0.057	24	0.152	**0.833**	-0.025	45	-0.037	**0.421**	0.709	66	-0.369	0.341	**0.775**
4	0.342	-0.103	**-0.643**	25	**0.455**	0.579	0.038	46	0.336	-0.257	**0.586**	67	0.106	0.008	**0.817**
5	0.193	-0.040	0.085	26	0.073	**0.749**	0.313	47	0.212	**0.830**	-0.046	68	-0.049	-0.338	**0.916**
6	0.375	**0.467**	-0.262	27	**0.941**	0.063	-0.483	48	0.259	**0.817**	-0.141	69	-0.651	0.245	**0.660**
7	-0.237	**0.466**	-0.391	28	**0.734**	0.375	-0.121	49	-0.037	**0.974**	0.023	70	**0.564**	0.242	-0.011
8	0.344	**0.480**	-0.187	29	**0.886**	0.128	-0.034	50	**0.625**	0.390	0.023	71	**0.682**	0.155	0.027
9	**0.688**	0.131	-0.537	30	**1.042**	-0.044	-0.099	51	0.088	**0.524**	-0.673	72	-0.098	0.469	**0.545**
10	**1.170**	-0.406	-0.077	31	**0.962**	-0.111	-0.025	52	-0.065	0.301	**0.617**	73	0.212	-0.071	**0.862**
11	**0.537**	0.117	-0.607	32	**0.956**	0.095	-0.195	53	0.389	-0.155	**0.515**	74	**0.583**	0.301	0.305
12	**0.960**	0.024	-0.204	33	**0.799**	0.072	0.167	54	0.022	**0.852**	0.174	75	**0.740**	0.149	0.254
13	**0.961**	-0.082	0.049	34	0.443	**0.573**	0.060	55	0.018	**0.951**	-0.095	76	0.053	**0.768**	0.259
14	**0.589**	0.456	-0.037	35	0.450	**0.469**	0.206	56	-0.157	**1.021**	-0.457	77	**0.575**	0.314	0.187
15	**0.900**	0.125	-0.178	36	0.391	**0.634**	0.053	57	-0.119	**0.815**	0.240	78	0.286	0.194	**0.654**
16	**0.940**	-0.391	0.343	37	0.935	**-0.618**	0.408	58	-0.245	0.398	0.336	79	-0.029	**0.903**	0.179
17	**0.467**	0.301	-0.350	38	0.294	**0.692**	0.057	59	-0.307	**0.903**	-0.067	80	**0.661**	-0.147	0.276
18	**0.502**	0.583	-0.069	39	0.400	**0.548**	0.239	60	-0.352	**0.900**	0.210	81	**0.492**	0.156	0.273
19	**1.001**	-0.255	0.008	40	0.304	**0.432**	0.453	61	-0.175	0.379	**0.638**	82	0.498	-0.047	**0.548**
20	**0.977**	-0.212	0.003	41	**0.792**	0.114	0.062	62	-0.148	**1.069**	-0.014	83	**0.474**	0.273	0.343
21	**0.724**	0.241	-0.244	42	**0.602**	0.091	-0.182	63	-0.163	**1.119**	-0.125				

Note. Factor pattern coefficients greater than |.4| are bolded.

Table 2. *Structure Coefficients*

Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3
1	0.794	0.587	0.565	22	0.804	0.820	0.321	43	0.776	0.740	0.471	64	0.725	0.509	0.808
2	0.263	0.445	0.192	23	0.812	0.801	0.418	44	0.737	0.696	0.798	65	0.898	0.779	0.398
3	0.365	0.446	0.155	24	0.661	0.919	0.364	45	0.530	0.673	0.856	66	0.176	0.412	0.748
4	0.001	-0.140	-0.536	25	0.832	0.877	0.458	46	0.428	0.180	0.631	67	0.463	0.391	0.866
5	0.205	0.114	0.153	26	0.675	0.917	0.636	47	0.709	0.944	0.367	68	0.135	-0.013	0.764
6	0.553	0.599	0.080	27	0.773	0.462	-0.053	48	0.707	0.923	0.287	69	-0.215	0.096	0.475
7	-0.115	0.167	-0.312	28	0.915	0.785	0.340	49	0.579	0.960	0.385	70	0.710	0.589	0.325
8	0.562	0.621	0.147	29	0.951	0.666	0.397	50	0.877	0.788	0.443	71	0.789	0.590	0.380
9	0.538	0.350	-0.191	30	0.973	0.567	0.333	51	0.125	0.318	-0.431	72	0.429	0.620	0.685
10	0.884	0.293	0.269	31	0.881	0.478	0.345	52	0.388	0.500	0.706	73	0.538	0.396	0.926
11	0.349	0.216	-0.331	32	0.932	0.615	0.253	53	0.514	0.288	0.622	74	0.902	0.783	0.673
12	0.887	0.543	0.218	33	0.915	0.634	0.538	54	0.627	0.934	0.514	75	0.942	0.709	0.630
13	0.931	0.536	0.431	34	0.825	0.872	0.473	55	0.569	0.925	0.282	76	0.643	0.902	0.580
14	0.857	0.809	0.394	35	0.831	0.829	0.582	56	0.282	0.746	-0.128	77	0.851	0.745	0.557
15	0.901	0.616	0.258	36	0.809	0.898	0.468	57	0.491	0.833	0.505	78	0.688	0.626	0.853
16	0.844	0.327	0.595	37	0.726	0.123	0.571	58	0.148	0.376	0.385	79	0.610	0.954	0.517
17	0.504	0.456	-0.032	38	0.750	0.897	0.453	59	0.226	0.686	0.151	80	0.689	0.372	0.503
18	0.835	0.869	0.373	39	0.843	0.889	0.623	60	0.299	0.762	0.408	81	0.706	0.568	0.545
19	0.846	0.372	0.340	40	0.768	0.797	0.752	61	0.335	0.518	0.710	82	0.704	0.476	0.744
20	0.846	0.398	0.341	41	0.890	0.631	0.446	62	0.512	0.972	0.338	83	0.792	0.701	0.653
21	0.769	0.597	0.161	42	0.580	0.395	0.112	63	0.479	0.968	0.239				

The Promax rotation produced correlated first-order factors, so these overlapped factors implied the existence of a higher order factor that could be extracted from the inter-factor correlation matrix. The second-order factor was extracted from the inter-factor correlation matrix. Table 3 contains the correlation matrix for Promax rotated first-order factors and the second order factor. The eigenvalue for the second-order factor was 1.969, and it explained 66% of the variance in the inter-factor correlation matrix.

Table 3. *Inter-Factor Correlation Matrix and Second-Order Factor Structure Coefficients*

	Correlation Matrix			Second-Order Coefficients
	First1	First2	First3	
First1	1			0.861
First2	0.623	1		0.842
First3	0.430	0.388	1	0.720

First-order factors were not observed variables; they were abstractions of observed variables. Because second-order factors were extracted from the inter-factor correlation matrix, second-order factors were “abstractions of abstractions” (Thomson, 2004, p. 74). We would not want to interpret second-order factors in terms of abstractions (e.g., first-order factors); instead, we should interpret second-order factors in terms of measured variables. To be able to interpret second-order factors in terms of measured variables, the Schimid and Leiman (1957) solution is provided in Table 4.

Table 4. *Schmid and Leiman Solution*

Item #	Second	First1	First2	First3
1	0.804	0.160	0.029	0.125
2	0.376	-0.008	0.132	0.014
3	0.407	0.041	0.107	-0.027
4	-0.255	0.088	-0.030	-0.310
5	0.194	0.050	-0.012	0.041
6	0.528	0.097	0.136	-0.126
7	-0.093	-0.061	0.136	-0.188
8	0.566	0.089	0.140	-0.090
9	0.316	0.178	0.038	-0.259
10	0.610	0.303	-0.118	-0.037
11	0.124	0.139	0.034	-0.292
12	0.700	0.248	0.007	-0.098
13	0.794	0.249	-0.024	0.024
14	0.865	0.152	0.133	-0.018
15	0.752	0.233	0.036	-0.086
16	0.727	0.243	-0.114	0.165
17	0.404	0.121	0.088	-0.169
18	0.874	0.130	0.170	-0.033
19	0.653	0.259	-0.074	0.004
20	0.665	0.253	-0.062	0.001
21	0.651	0.187	0.070	-0.118
22	0.819	0.133	0.158	-0.053
23	0.850	0.131	0.139	0.006
24	0.814	0.039	0.242	-0.012
25	0.907	0.118	0.168	0.018
26	0.919	0.019	0.218	0.151
27	0.516	0.243	0.018	-0.233
28	0.861	0.190	0.109	-0.058
29	0.846	0.229	0.037	-0.016
30	0.789	0.269	-0.013	-0.048
31	0.717	0.249	-0.032	-0.012
32	0.763	0.247	0.028	-0.094
33	0.869	0.207	0.021	0.080
34	0.907	0.115	0.167	0.029
35	0.931	0.116	0.136	0.099
36	0.909	0.101	0.184	0.026
37	0.578	0.242	-0.180	0.197
38	0.877	0.076	0.201	0.027
39	0.978	0.103	0.159	0.115
40	0.952	0.079	0.126	0.218
41	0.823	0.205	0.033	0.030
42	0.464	0.156	0.026	-0.088
43	0.828	0.124	0.116	0.053
44	0.912	0.085	0.081	0.264
45	0.833	-0.010	0.122	0.342
46	0.495	0.087	-0.075	0.282
47	0.848	0.055	0.241	-0.022
48	0.809	0.067	0.238	-0.068
49	0.805	-0.010	0.283	0.011
50	0.883	0.162	0.113	0.011
51	0.033	0.023	0.152	-0.324
52	0.642	-0.017	0.088	0.297
53	0.575	0.101	-0.045	0.248
54	0.862	0.006	0.248	0.084
55	0.748	0.005	0.277	-0.046
56	0.396	-0.041	0.297	-0.220
57	0.757	-0.031	0.237	0.116
58	0.366	-0.063	0.116	0.162
59	0.448	-0.079	0.263	-0.032
60	0.606	-0.091	0.262	0.101
61	0.628	-0.045	0.110	0.307
62	0.763	-0.038	0.311	-0.007
63	0.712	-0.042	0.326	-0.060
64	0.831	0.123	-0.007	0.295
65	0.873	0.178	0.106	-0.019
66	0.527	-0.095	0.099	0.373
67	0.686	0.027	0.002	0.394
68	0.333	-0.013	-0.098	0.441
69	0.121	-0.168	0.071	0.318
70	0.681	0.146	0.070	-0.005
71	0.737	0.176	0.045	0.013
72	0.703	-0.025	0.136	0.263
73	0.743	0.055	-0.021	0.415
74	0.975	0.151	0.088	0.147
75	0.945	0.191	0.043	0.122
76	0.879	0.014	0.223	0.125
77	0.894	0.149	0.091	0.090
78	0.880	0.074	0.056	0.315
79	0.864	-0.008	0.263	0.086
80	0.644	0.171	-0.043	0.133
81	0.752	0.127	0.045	0.132
82	0.784	0.129	-0.014	0.264
83	0.885	0.123	0.079	0.165

Note. Factor pattern coefficients lower than |.4| are underlined.

Separate EFAs for each category

Reliability structure. Internal consistency reliability estimates for each category were calculated. The reliability estimate for the first Category, which had 32 items, was 0.978. Cronbach's α for the second Category, which included 24 items, was 0.939, and the estimate for the third Category, which had 27 items, was 0.969. Although the third Category's reliability estimate was 0.939, it was very close to 1. The other two categories' estimates were close to 1; therefore, we could conclude that our scores were reliable for each category.

Factor structure. Table 5 shows the pattern/structure coefficients for each category. Three separate EFAs were conducted by using the principal components method. The first EFA for Category 1 with 32 items resulted in only 5 pattern/structure coefficients smaller than .4. A second separate EFA for Category 2 with 24 items (33-56) resulted in only 1 item (51) and had a value smaller than .4. The other separate EFA for Category 3 with 27 items produced only 1 item and had a value smaller than .4. All others had pattern/structure coefficients bigger than .4 absolute value.

Table 5. *Pattern/Structure Coefficient for Each Category*

Category 1		Category 2		Category 3	
Item #	Factor 1.1	Item #	Factor 2.1	Item #	Factor 3.1
1	**0.792**	33	**0.869**	57	**0.651**
2	0.357	34	**0.963**	58	**0.558**
3	**0.432**	35	**0.985**	59	**0.534**
4	0.013	36	**0.975**	60	**0.594**
5	0.273	37	**0.489**	61	**0.719**
6	**0.636**	38	**0.908**	62	**0.741**
7	0.025	39	**0.963**	63	**0.704**
8	**0.683**	40	**0.956**	64	**0.841**
9	**0.611**	41	**0.875**	65	**0.815**
10	**0.817**	42	**0.546**	66	**0.626**
11	0.353	43	**0.886**	67	**0.735**
12	**0.846**	44	**0.898**	68	**0.411**
13	**0.929**	45	**0.797**	69	0.146
14	**0.923**	46	**0.467**	70	**0.791**
15	**0.914**	47	**0.900**	71	**0.816**
16	**0.758**	48	**0.884**	72	**0.815**
17	**0.524**	49	**0.844**	73	**0.810**
18	**0.904**	50	**0.864**	74	**0.933**
19	**0.768**	51	0.155	75	**0.901**
20	**0.824**	52	**0.671**	76	**0.895**
21	**0.829**	53	**0.598**	77	**0.910**
22	**0.902**	54	**0.897**	78	**0.895**
23	**0.852**	55	**0.825**	79	**0.793**
24	**0.772**	56	**0.511**	80	**0.709**

(continued)

Item #	Factor 1.1	Item #	Factor 2.1	Item #	Factor 3.1
25	**0.889**			81	**0.833**
26	**0.759**			82	**0.870**
27	**0.800**			83	**0.946**
28	**0.950**				
29	**0.963**				
30	**0.931**				
31	**0.888**				
32	**0.915**				

Note. Factor pattern coefficients greater than |.4| are bolded.

The extracted factor from the 32 by 32 inter-variable correlation matrix for Category 1, explained around 58 % of variation. The other one factor extracted from 24 items for Category 1 explained 65 % of the variation. The factor extracted from 27 items inter-variable correlation matrix for Category 3, explained 58 % of the variation among variables.

Discussion

The four phases of the study shed important light on the usefulness of the AMLE's SIA. The validity indicators are reasonable. The benchmarks by which the validly was judged are not the only ones available but do comprise prominent means for determining that the instrument does measure what it is intended to measure. Arguably, the strongest indicator for validity claims comes from the ratings of experts as compared to novices unfamiliar with the AMLE SIA or the principles on which it is based. The novice group was barely above chance score indicating that guessing could likely have accounted for the same score while the experts were much closer to par score. It is very important to note that the experts were able to pick the distractors out 100% of the time. The misclassification of items as not being representative of *This We Believe. . .* was a strong indicator for further analysis and instrument refinement.

The refinement of the instrument led to experts having to sit around a table, listen to each other explain their rationale for their ratings and to come to consensus on what to do with particular items. The analysis process reduced the number of items and increased the clarification of items that previously caused ambiguity.

Perhaps the most detrimental aspect to validity and reliability is instrument length. In the original version items were presented within characteristic and category but one item at a time. The process required 96 mouse clicks to progress through the instrument with additional clicks for consent, information, and terminology. When looking at the graphs there was not explanation for why as the items progressed it would take respondents more time to answer an item. So the items were randomized to determine if the items toward the end of the assessment were more time consuming. In fact, they were not and time to completion increased. We reduced the total number of mouse clicks to 16 with the same number of introductory clicks. Presenting a grouping of items by characteristic within the category reduced the amount of reading and allowed respondents to simply move a slider. They were able to see which items came before and after the item they were addressing. The perception could have been that there were fewer items. In fact, the time to completion of the instrument was reduced to ¼ the time needed previously and the reliability was very high at .98. Therefore, we can rule out that the respondents simply guessed without reading the prompts.

Finally, the EFA clearly indicates that the dropped items and the revised items provide a reasonable fit to the data, and the new instrument accounted for 73% of the variance, a very respectable number. The factor loadings were mostly as expected with a few items being problematic. Some items that we expected to load on Category 3 actually loaded on Category 1. The potential problem here is that respondents may have viewed the culture and community items as they relate to curriculum because the items do have implications for instruction.

The revised instrument does yield valid and reliable scores. However, more data is necessary for a fine-grained analysis. That data also needs to be more representative of middle school programs and the way the middle school concept is enacted in the U.S. While we believe the analysis is robust, the sample size for the EFA only accounts for a single school district in a single state. The results seem to indicate that fewer items could improve the instrument, as could the collapsing of some of the characteristics.

References

Bryant, F. (2000). Assessing the validity of measurement. In L. G. Grimm, & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 99-139). Washington, D.C.: American Psychological Association.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Cook, C., Faulkner, S., & Kinne, L. (2009). Indicators of middle school implementation: How do Kentucky's schools to watch measure up? *Research in Middle Level Education Online, 32*, 1-10.

Cook, C., Heath, F., Thompson, R. L., & Thompson, B. (2001). Score reliability in webor internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, *61*(4), 697-706.

de Winter, J. C. F., Dodou, D. & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*(2), 147-181.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549-576.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.

Authors (2004).

Horn, J. L. (1965). A rationale and test for the number if factors in factor analysis. *Psychometrika*, *30*, 179-185.

Jackson, S., & Lunenburg, F. (2010). School performance indicators, accountability ratings, and student achievement. *American Secondary Education, 39*, 27-44.

McEwin, C. K., & Greene, M. W. (2011). *The status of programs and practices in America's middle schools: Results from two national studies*. Association for Middle Level Education. Retrieved from http://www.amle.org/portals/0/pdf/research/Research_from_the_Field/Status_Programs_Practices_AMLE.pdf.

Mertens, S. B., & Flowers, N. (2006). Middle Start's impact on comprehensive middle school reform. *Middle Grades Research Journal, 1*, 1-26.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*, 159-168.

National Middle School Association. (2010). *This we believe: Keys to educating young adolescents*. Westerville, OH: National Middle School Association.

National Middle School Association. (2014). *This we believe: Keys to educating young adolescents position paper of national middle school association executive summary*. Retrieved from http://www.uww.edu/Documents/colleges/coeps/academics/This_We_Believe_Exec_Summary.pdf

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavioral Research Methods, Instrument, & Computers*, *32*(3), 396-402.

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics, 32*, 153-161.

Schimid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrica*, *22*, 53-61.

Thompson, B. (1992). A partial test distribution for cosines among factors across samples, In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp. 81-97). Greenwich, CT: JAI Press.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical review and some guidelines. *Educational and Psychological Measurement, 56*(2), 197-208.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Trochim, W. M. K. (2006). *Measurement validity types*. Research Methods Knowledge Base. Retrieved from http://www.socialresearchmethods.net/kb/measval.php

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher*, *38*(5), 343-352.